

Deep Learning in Drug Discovery and Development

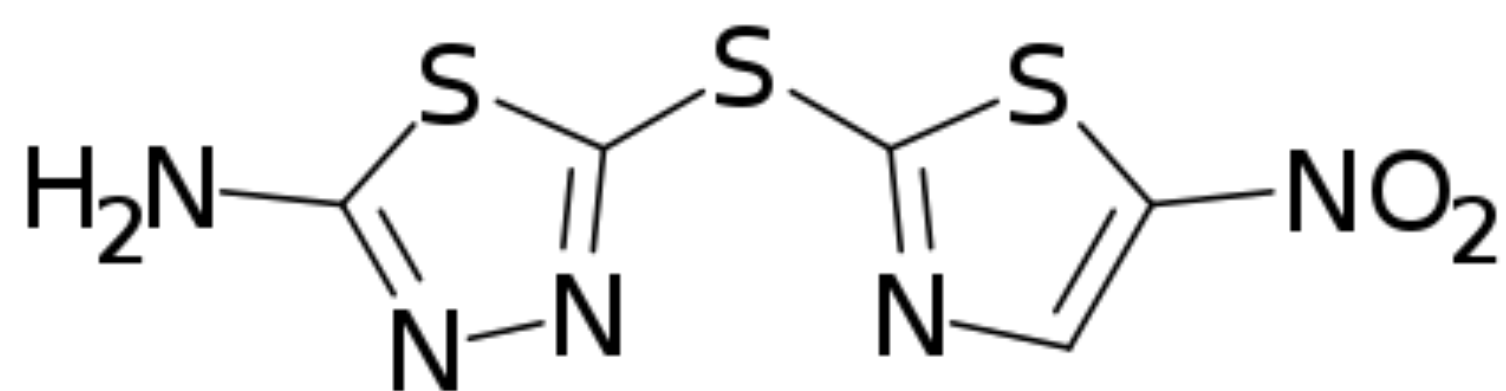
Khoa Nguyen

kanguyen18@earlham.edu

Department of Data Science, Earlham College

1. Introduction

- Drug discovery and development is a costly and time-consuming process, taking up to billions of dollars and around 12-15 years.
- In 2020, researchers at MIT discovered a new powerful antibiotic compound with an entirely new mechanism, named Halicin, using deep learning (DL) with graph neural networks (GNN).



Halicin chemical structure

- After that, they continue to identify 23 candidates that were structurally dissimilar from existing antibiotics within 3 days.
- This project discusses the feasibility of some state-of-the-art GNNs in predicting solubility of commercially available compounds in the ZINC database.

2. Data Diagram

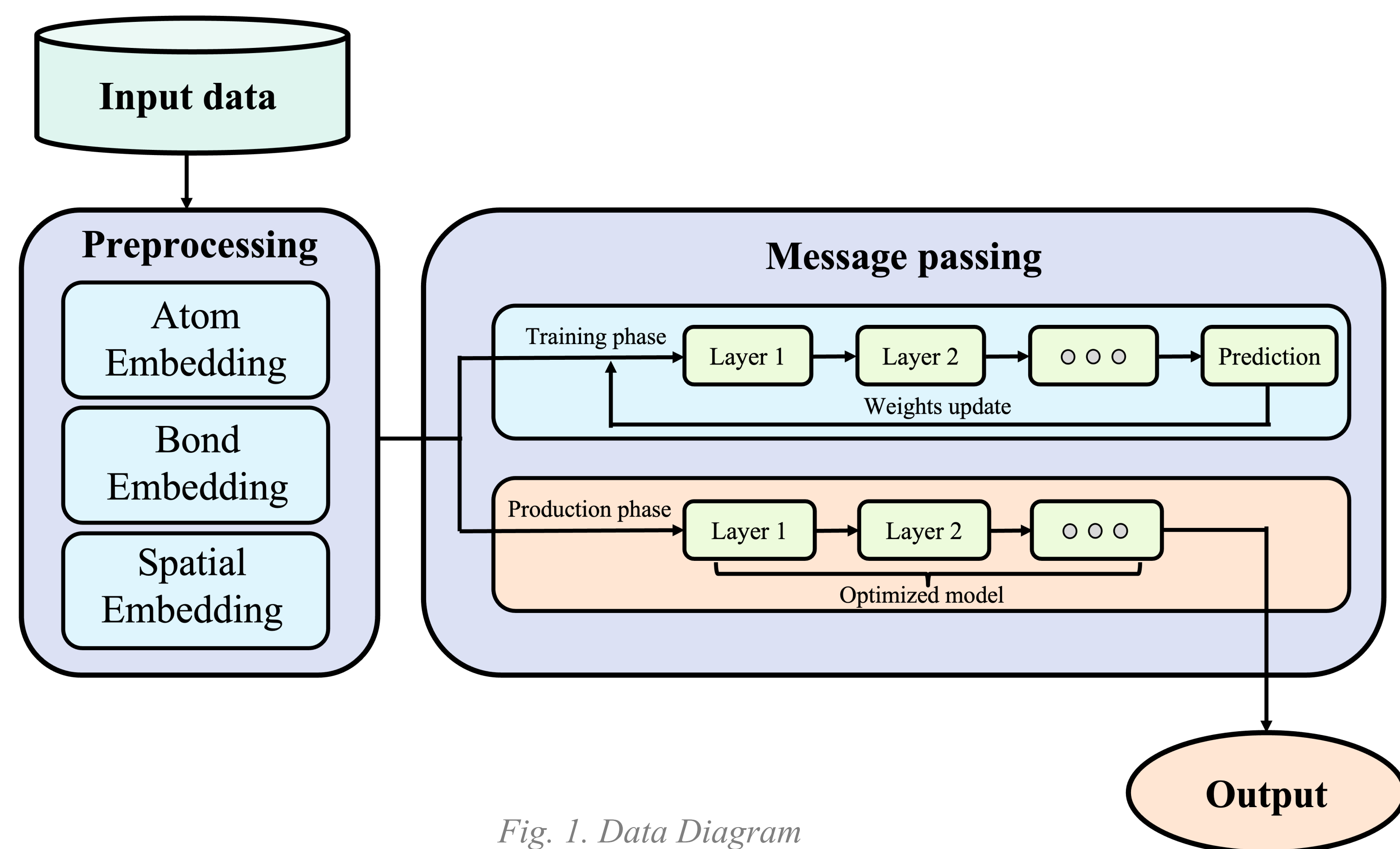


Fig. 1. Data Diagram

3. Methodology

1. Dataset

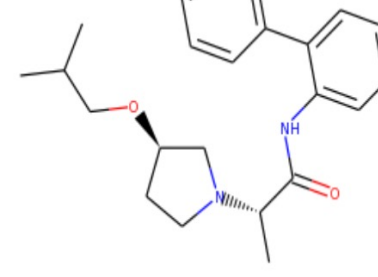
- ZINC contains over 230 million purchasable compounds in SMILES, 2D, and 3D representation.

ZINC19813322

In: [anodyne](#) [for-sale](#) [on-demand](#)

[Google](#) [Pubmed](#) [PubMed](#)

Added	Availability	Since	Mwt	logP	Download
2015-08-08	On-Demand	2015-08-08	386.505	4.428	↓



Draw

Mol Formula	Rings	Heavy Atoms	Hetero Atoms	Fraction sp ³	Tranche
C23H30N2O2	3	27	4	0.43	FIAD

SMILES	<chem>C[C@C](O)[C@H](1CCN1)[C@H](C)[C@@H](O)C(=O)N(C)C(=O)C2CCCC2-C2CCCC2C1</chem>	↓
InChI	InChI=1S/C23H30N2O2/c1-17/2-16-27-20-13-14-25/(5-10/16/32/26-24-22-12-8-7-11-21/22/19-9-5-4-6-10-19/4-1)	↓
InChI Key	DUTOSIREQGECD-AZUAAIDMISA-N	↓

Available 3D Representations

pH range	Net charge	H-bond donors	H-bond acceptors	IPSA	Rotatable bonds	Apolar desolvation	Polar desolvation	Download
Reference	0	1	3	41	7	9.67	-8.55	↓
Mid pH (7.4)	1	2	2	42	7	12.06	-36.61	↓
Mid pH (7.4)	1	2	2	42	7	11.95	-36.73	↓

An instance of purchasable compounds in ZINC database

- logP is the partition coefficient, measuring the solubility of a chemical substance.

2. Graph neural networks

- Message-passing neural networks (MPNN): Each node in the graph has a hidden state that gets updated based on its previous state and new messages from neighboring nodes.
- In this project, the message-passing GNNs are GCN, Gated-GCN, and GAT.

4. Implementation

Input: subset of ZINC database with 10000 graphs for training, 1000 for validation, and 1000 for testing.

Models: GAT, Gated-GCN, GCN, and MLP with approximately 100.000 weight parameters each.

Hyperparameters:

- Trained using mini batched gradient descent with batch size of 128 and 1000 epochs.
- Initial learning rate is 0.001 with reduce factor of 0.5 conditioning on schedule patient of 10.
- Early stopping happens at minimum learning rate of 1e-5.

Performance metric: Mean Absolute Error score (MAE). Smaller MAE means better performance.

Hardware: Earlham College’s Layout cluster.

5. Results

The test MAE scores for each model were recorded after each epoch run and visualized with Tensorboard (*Fig. 2a and 2b*):

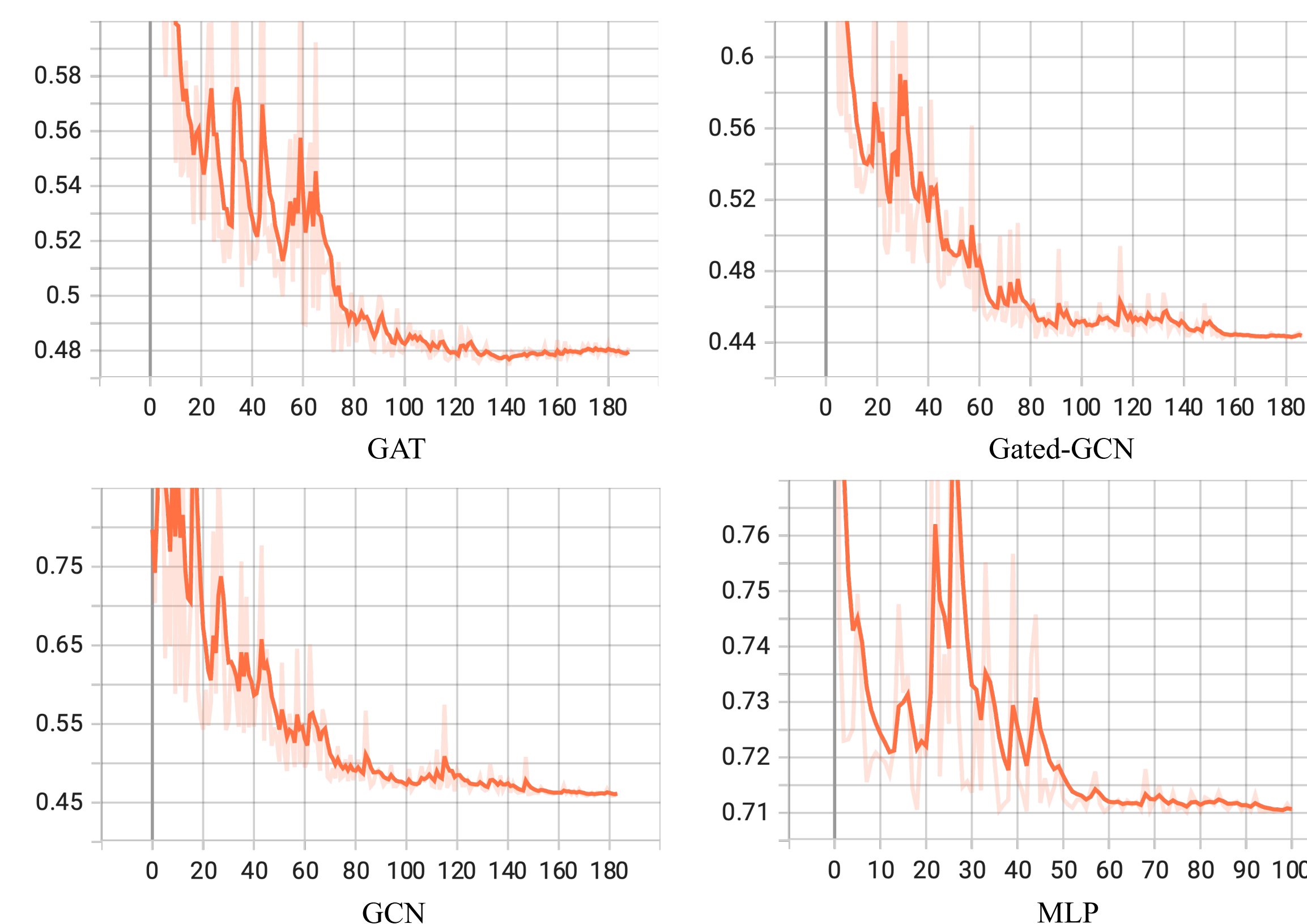


Fig. 2a. Performance on test sets of 4 GNNs. Vertical axis is MAE score. Horizontal axis is the number of epochs.

Model	Test MAE	Train MAE	Total epochs	Total time (hrs)
GAT	0.4811	0.2953	188	0.8087
Gated-GCN	0.4428	0.2868	186	0.4978
GCN	0.4629	0.3543	183	0.2518
MLP	0.7103	0.6445	100	0.0828

Fig. 2b. Training summary

6. Discussion

- For future work, I will add newly designed architectures into the benchmarking GNNs repository. Other molecular datasets like QM9, tox21, HIV, etc. can also be added.
- The current implementation only utilized a small subset of the ZINC database (12k compounds) and less computational expensive architectures with around 100k weight parameters for each model. In addition to that, the training process only used CPU and did not take full advantage of available hardware like GPUs and parallel computing. These can be optimized for faster training time and more accurate results.
- The discussion of this project only reflected a small subset of computational drug discovery and development. Current industry practices have already integrated deep learning into most phrases of drug R&D.

7. Acknowledgements

I would like to thank Dr. David Barbella, Dr. Charlie Peck, Dr. Michael Lerner, Porter Libby, and Khai Nguyen for all the supports.